# Chapter 1
# Technologies for Big Data

**Kapil Bakshi**
*Cisco Systems Inc., USA*

## ABSTRACT

*This chapter provides a review and analysis of several key Big Data technologies. Currently, there are many Big Data technologies in development and implementation; hence, a comprehensive review of all of these technologies is beyond the scope of this chapter. This chapter focuses on the most popularly accepted technologies. The key Big Data technologies to be discussed include: Map-Reduce, NOSQL technology, MPP (Massively Parallel Processing), and In Memory Databases technologies. For each of these Big Data technologies, the following subtopics are discussed: the history and genesis of the Big Data technologies, problem set that this technology solves for Big Data analytics, the details of the technologies, including components, technical architecture, and theory of operations. This is followed by technical operation and infrastructure (compute, storage, and network), design considerations, and performance benchmarks. Finally, this chapter provides an integrated approach to the above-mentioned Big Data technologies.*

## INTRODUCTION: THE CHALLENGE OF BIG DATA

The amount of data in the world is being collected and stored at unprecedented rates. A study by IDC Gantz & Reinsel, (2011) indicates that the world's information is doubling every two years. Also the IDC study by Gantz & Reinsel (2011), mentions that the world created a staggering 1.8 zettabytes of information (a zettabyte is 1000 exabytes), and projections suggest that by 2020, we'll generate will generate 50 times that amount.

Big Data has been defined as, when data sets get so large, that traditional technologies,

DOI: 10.4018/978-1-4666-4699-5.ch001

techniques, and tools for extracting insights are no longer useful in a reasonable timeframe and cost-effective manner. This has spawned a new generation of technologies and corresponding considerations. Desai, Kommu & Rapp (2011) examine the cause of this explosion of Big Data, the following factors dominate:

- **Mobility trends:** Mobile devices and sensor proliferation;
- **New data access:** Internet, interconnected systems, and social networking;
- **Open source model:** Major changes in the information processing model and the availability of an open source framework.

What distinguishes Big Data from data in the past, however, is not just its vast volume. The defining features of Big Data are also its variety—the sources and types of data being collected—and its velocity, the speed at which the data is flowing through the networked systems. Studies like Cisco Virtual Networking Index by Barnett, (2011) estimate that in 2016, global IP traffic will reach 1.3 zettabytes per year or 110.3 exabytes per month. Moreover, it is anticipated that there will be 19 billion networked devices by 2016.

One of the most interesting aspects about Big Data is that that unstructured data is the fastest growing type of data. Unstructured data refers to information that either does not have a predefined data model or does not fit well into relational database tables. Examples of unstructured data include imagery, sensor data, telemetry data, video, documents, log files, and email files. The challenge is not only to store and manage this vast mix, but to analyze and extract meaningful value from it—and to do so in a reasonable timeframe and at a reasonable cost.

Fortunately, a new generation of technologies has emerged for collecting, storing, processing, and analyzing Big Data. This chapter provides a survey of these technologies, including implementation and operational details:

- MapReduce framework, including Hadoop Distributed File System (HDFS);
- NoSQL (Not Only SQL) data stores;
- MPP (Massively Parallel Processing) databases;
- In-memory database processing.

## MAPREDUCE AND HADOOP DISTRIBUTED FILE SYSTEM

In 2003, Google published a paper on Google File Systems (GFS) by Ghemawat, Gobioff & Leung (2003). and a subsequent paper on the MapReduce Dean & Ghemawat (2004) model to address processing large unstructured data sets. Hadoop, an open source framework developed by the Apache Foundation, is an outgrowth of the concepts presented by Google in these papers. The Hadoop project has both a distributed file system, called HDFS (Hadoop Distributed File System) modeled on GFS and a distributed processing framework using MapReduce concepts. This chapter will review Big Data Analytics technologies with the Hadoop project as a backdrop.

To get a big-picture understanding of the technology innovations embodied in Hadoop, let's start with the MapReduce programming model. The model, whose name comes from the map and reduces functions, uses a large number of computer nodes, connected via network interconnect, in a cluster fashion, to perform parallel processing across huge data sets.

Four characteristics —parallelism, fault tolerance, scalability, and data locality—are, in fact, the defining features of MapReduce systems, by White (2012):

- **Parallelism:** Breaking the large data sets into smaller compute and storage units makes it possible to perform analytics in parallel and therefore more efficiently. In addition, mappers and reducers do not communicate individually, so they can run